Generalised linear models for prediction: some principles, some programs and some practice

Nicholas J. Cox

University of Durham

n.j.cox@durham.ac.uk

Synopsis 1

Despite a history now over 30 years long, the adoption of generalised linear models (GLMs) remains patchy: they are well-known in several fields, but used little if at all in many others. One major advantage of GLMs is that they return predictions on the scale of the response. The use of link functions avoids the need for prior transformation of the response for back-transformation of predictions, and above all for bias corrections to back-transformations, whether systematic or *ad hoc*.

Synopsis 2

Case studies from environmental applications (suspended sediment concentrations of rivers, heights of forest trees) are introduced in which predictions on the response scale are of paramount scientific and practical interest. Heavy use is made of a suite of Stata programs written by the author producing graphic and numeric diagnostics after regression-type models, which extend and complement commands in official Stata. Most of these programs have uses beyond GLMs and they will also be discussed directly.

Generalised linear models 1

In what ways are these generalised?

The response distribution is from some **exponential family**. It could be, but need not be, normal. Its variance may be a function of the mean.

A so-called **link function** of the mean response μ is taken to be a linear function of predictors, i.e.
$\text{link}(\mu) = \beta_1 X_1 + \cdots + \beta_k X_k = \mathbf{X}\beta$.

The link function g must be monotone (invertible) and differentiable. Examples are identity $g(\mu) = \mu$, logarithm $\ln \mu$, reciprocal $1/\mu$.

E.g. model for C with logarithmic link is for $\ln(E(C))$; contrast to model with transformed response, which is for $E(\ln C)$.

Generalised linear models 2

To get predicted values of response, we estimate β by maximum likelihood and then invert the link to get $g^{-1}(\mathbf{X}\hat{\beta})$. If g is ln, g^{-1} is exp.

Loosely, a link function (other than identity) plays a role like a transformation of the response variable, but results are always produced and presented on the scale of the response. There is no need for back-transformation, bias corrections, etc.

It is still entirely possible to use, among the predictors $\mathbf{X}$, variables which are on transformed scales, e.g. $\ln Q$ or $1/Q$.

Generalised linear models 3

`glmcorr` (on SSC) calculates
correlation between response and fitted
and also RMSE.

Zheng and Agresti (*Statistics in Medicine* 2000) discuss this correlation as a general measure for GLMs.

Advantages:

- ◇ refers to response scale
- ◇ applicable to all types of GLM
- ◇ invariant under location-scale transformation
- ◇ root of fraction of variance explained

Limitations:

- ◇ need not match other definitions of R^2
- ◇ necessarily sensitive to outliers
- ◇ for different models and same data
- ◇ biased upwards (better jackknifed)

Example 1: Sediment concentrations

We seek, at a river gauging station, to relate

suspended sediment concentration C

to **discharge** (flux of water) Q,

(sometimes) **time** of measurement T
(a handle for hysteresis, seasonality, etc.)

and (rarely) other variables
(e.g. sediment supply).

We also often want to compare stations or look at long-term changes.

$C = C(Q)$ is known as a rating curve.

The problem could be approached from physical principles, but apart from several other difficulties we usually lack data on sediment supply.

Example 1: existing practice

Generalised linear models offer a systematic alternative to the **transformation, linear regression and fudge factor** approach which appears to be the most usual current practice.

Most suspended sediment rating curves are power functions $C = aQ^b$ fitted after logarithmic transformation by standard linear regression.

This corresponds to a model
$C = \exp(b_0 + b_1 \ln Q)$
in which the error is multiplicative and lognormal.
(Notation switch: $b_0 \leftarrow \ln a, b_1 \leftarrow b$.)

Statistical questions arising

◇ Back-transforming predictions to get $\exp(\widehat{\ln C})$ does not give unbiased predictions of C. The easiest ways to fix this are

1. to get variance of residuals s^2 and multiply by $\exp(s^2/2)$ (here called lognormal correction)
2. to get individual residuals e and multiply by mean of $\exp(e)$ (example of **smearing**).

◇ Which error distributions are appropriate?

◇ Alternative functional forms may be superior.

◇ Time series aspect is ignored: problems with autocorrelation of errors, alternative models possible.

◇ Assumes that Q is measured without error.

Residual plots

In examining any model, it is useful to look at a variety of extra special plots.

Official Stata supplies commands originally written for use after `regress`: `avplot` and `avplots`, `cprplot` and `acprplot`, `lvr2plot`, `rvfplot` and `rvpplot`. In September 2001, all but the first two were generalised to work after `anova`.

This suite omits some very useful kinds of plot. None of the commands may be used after other modelling commands. A new set of commands (on SSC) is designed for prediction of continuous responses. Updating for Stata 8 is in progress and programs will then be combined in one package (tentative name `modeldiag`).

Principles behind this package

◇ as far as possible, the command name by itself should produce a useful plot

◇ `predict` is used to produce temporary variables for residuals, fitted values, etc.

◇ each graph refers to the last model fitted

◇ each graph has reasonably smart default axis titles, etc.

◇ options are provided for key needs, e.g. `lowess` smoothing

(Note in contrast that many new graphics commands in Stata 8 fit models on the fly, but mostly plot observed and fitted against one covariate.)

Commands published so far

`anovaplot` (e.g. interaction plots)

`indexplot`

`ovfplot` (observed vs fitted)

`qfrplot` (quantile plots of fitted − mean and residuals)

`ofrtplot` (observed, fitted and residual vs time)

`rdplot` (residual distributions)

`regplot` (data and fitted vs first or named covariate)

`rvfplot2` (generalises `rvfplot`)

`rvlrplot` (residual vs lagged residual)

`rvpplot2` (generalises `rvpplot`)

Example 2: Tree heights

Tree height is not only of scientific interest to foresters but also a key variable in estimating timber (lumber) yield.

Predicting tree height from reflectance measures is needed to link field and satellite data.

Only positive heights make biological and practical sense, which can be ensured by using a log transformation. But back-transforming gives a biased estimate of height.

Once again, generalised linear models offer a systematic alternative.

Galloway: some numbers

Smearing correction	1.112
Lognormal correction	1.129

Model	Error	RMSE (m)	R
$\exp(b_0 + b_1\texttt{ref})$	normal	0.984	0.904
$\exp(b_0 + b_1\texttt{ref})$	gamma	1.000	0.905

`ref` is Landsat reflectance band 7

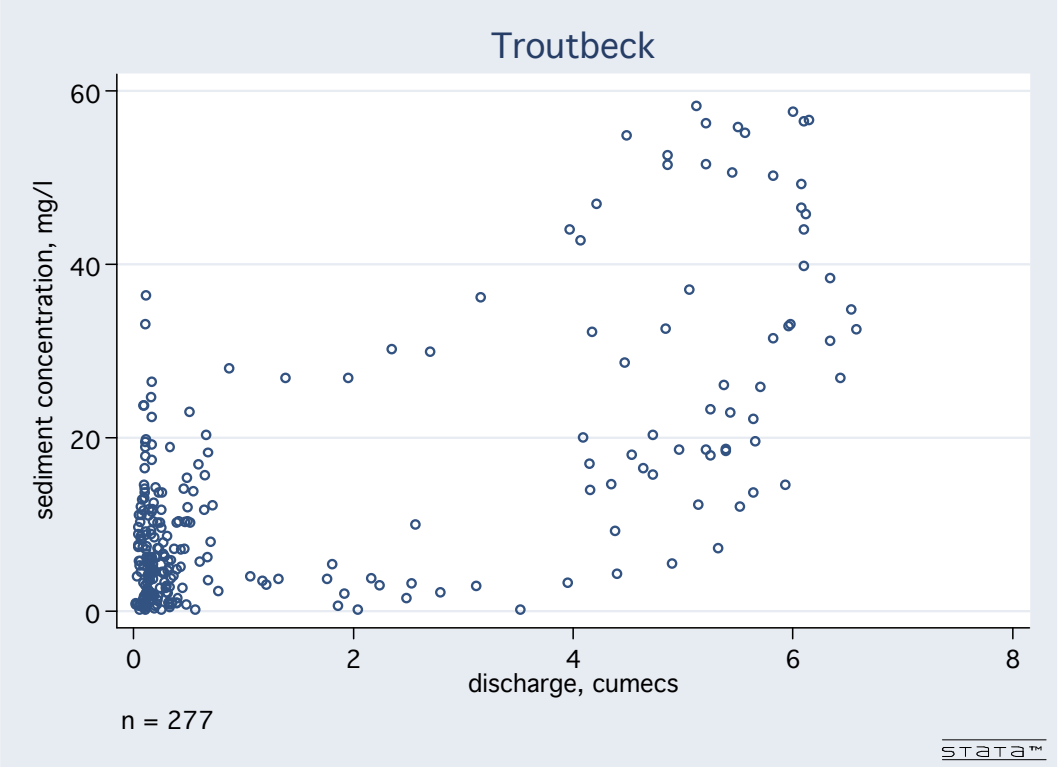
R is correlation(height, predicted height)

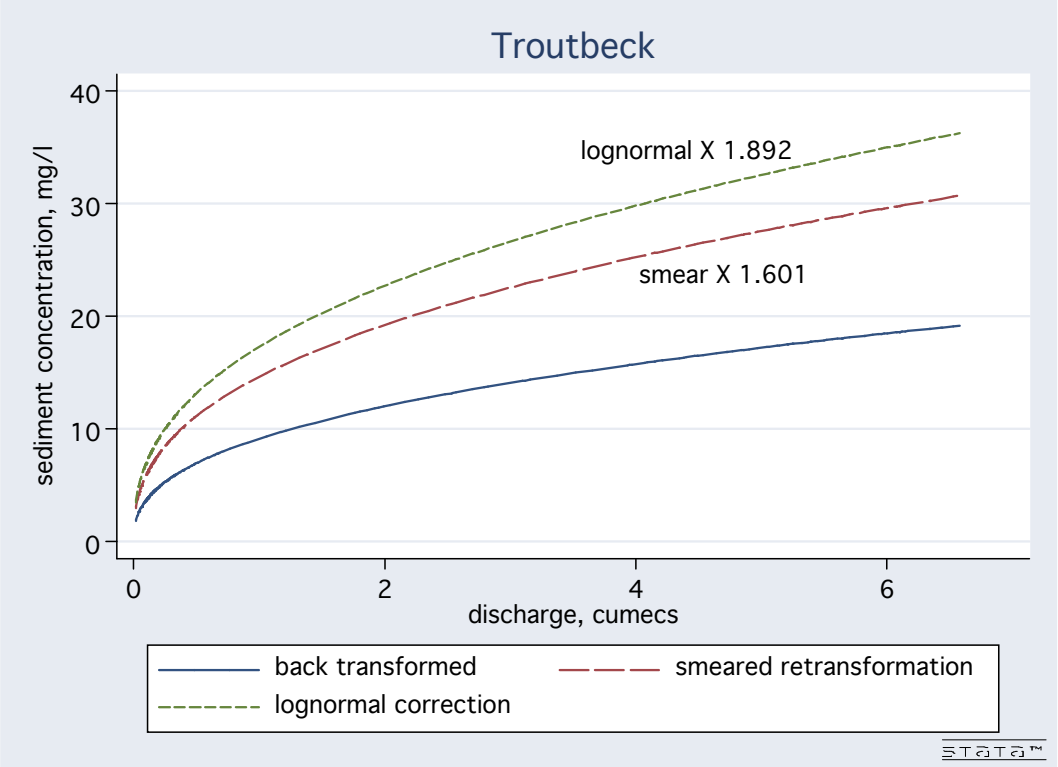
Troutbeck: some numbers

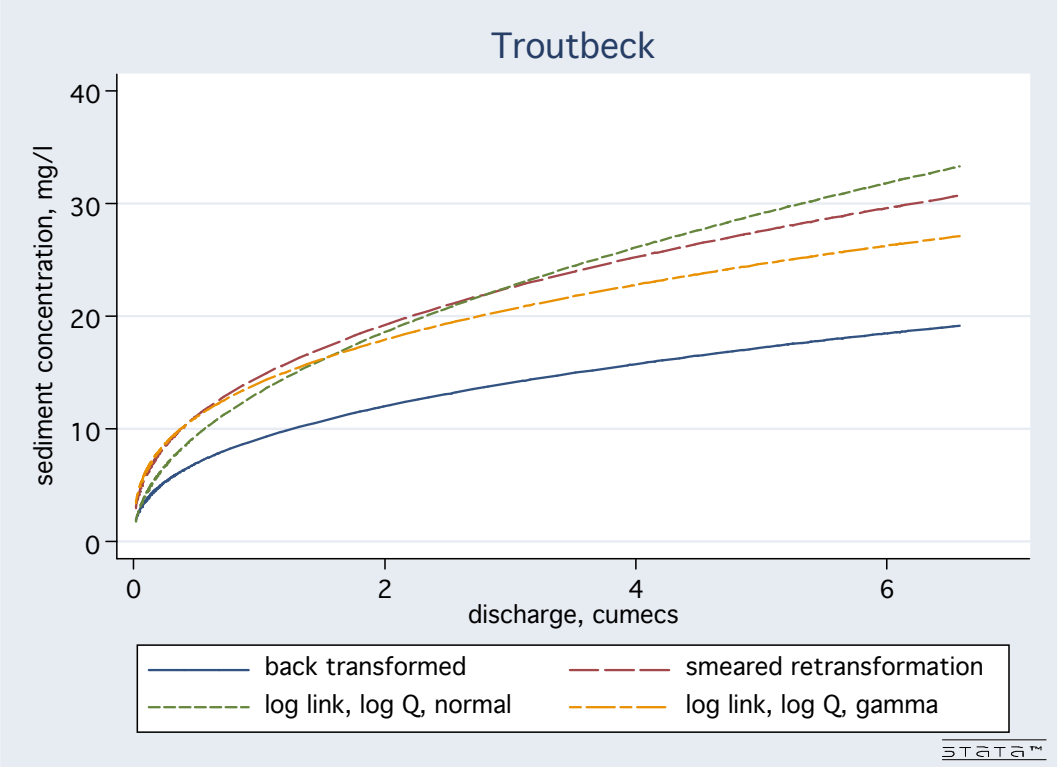
Smearing correction	1.691
Lognormal correction	1.892

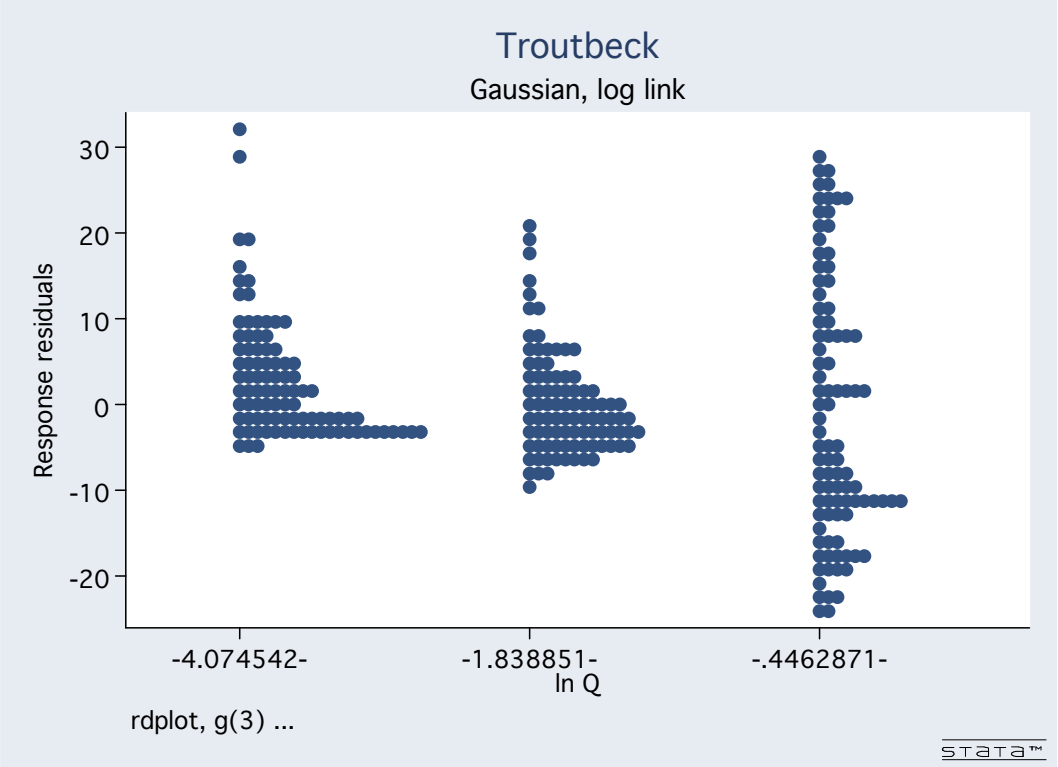
Model	Error	RMSE (mg/l)	R
$\exp(b_0 + b_1 \ln Q)$	normal	10.344	0.680
$\exp(b_0 + b_1 \ln Q)$	gamma	10.644	0.666
$\exp(b_0 + b_1 Q)$	normal	9.759	0.720
$\exp(b_0 + b_1 Q)$	gamma	9.760	0.720

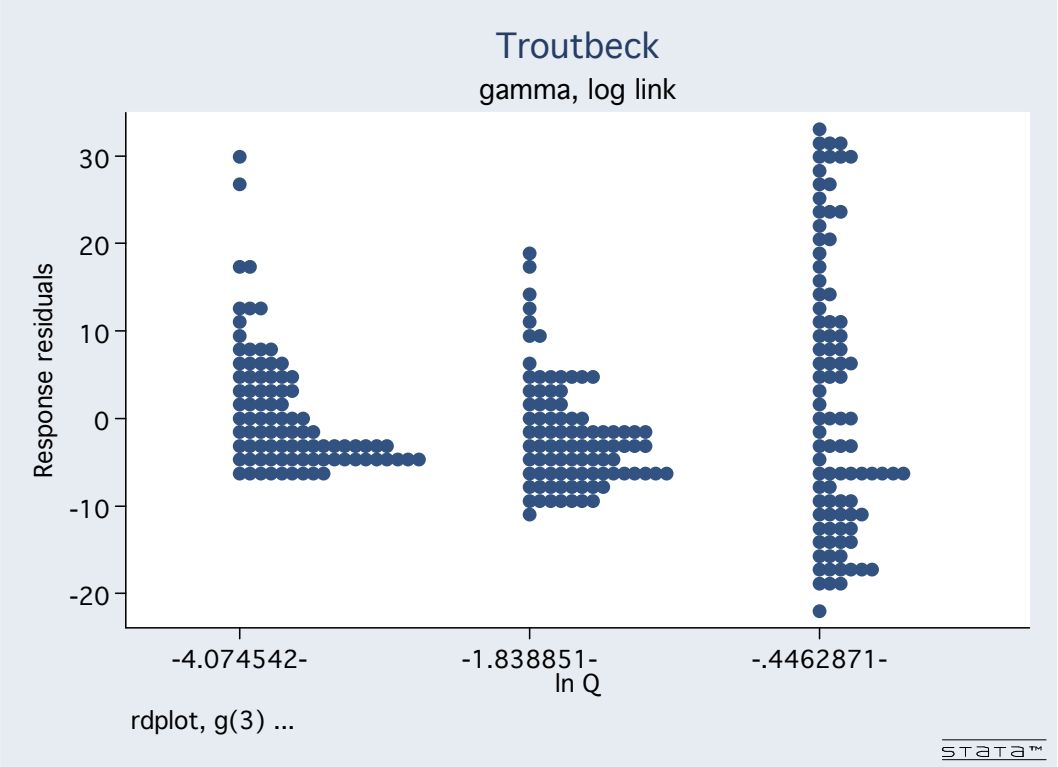
R is correlation$(C, \hat{C})$

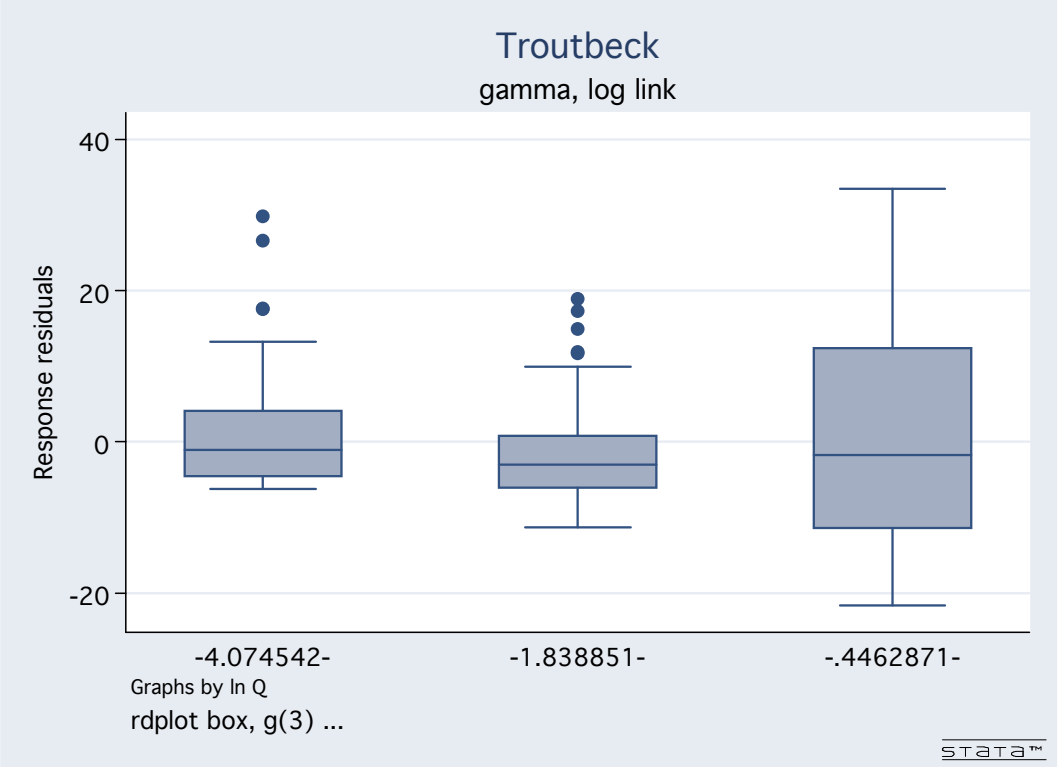

Troutbeck
sediment concentration, mg/l
0
20
40
60
0
2
4
6
8
discharge, cumecs
n = 277
stata™

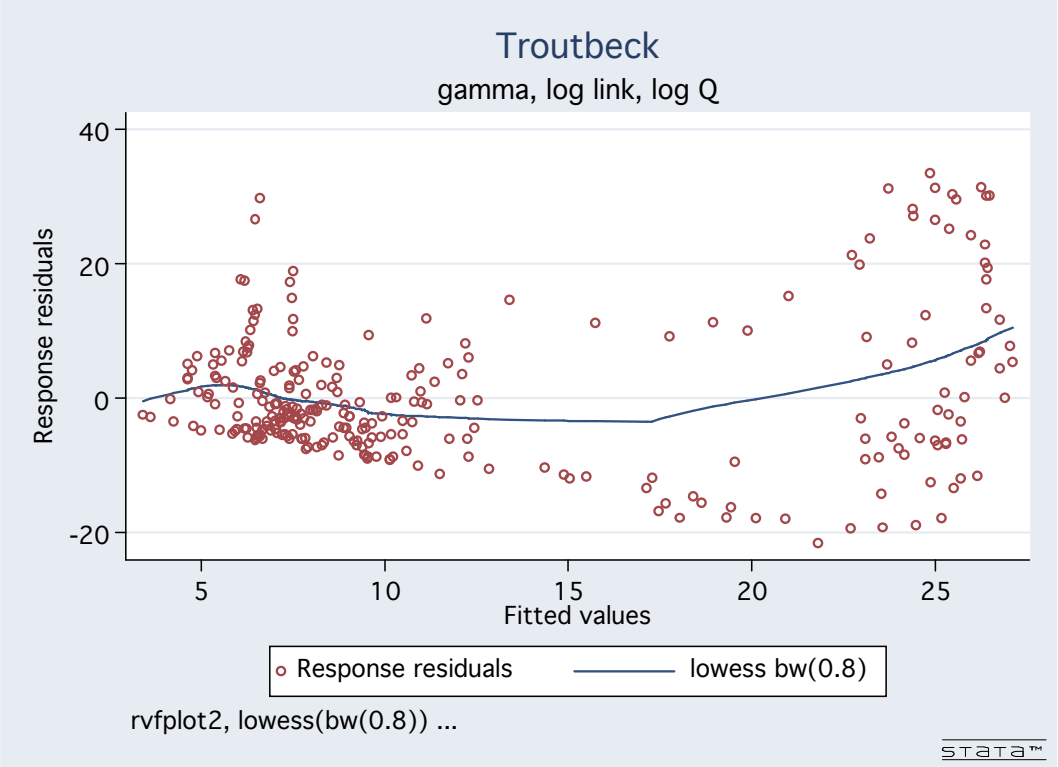

Troutbeck
sediment concentration, mg/l
40
30
20
10
0
lognormal X 1.892
smear X 1.601
0
2
4
6
discharge, cumecs
back transformed
smeared retransformation
lognormal correction

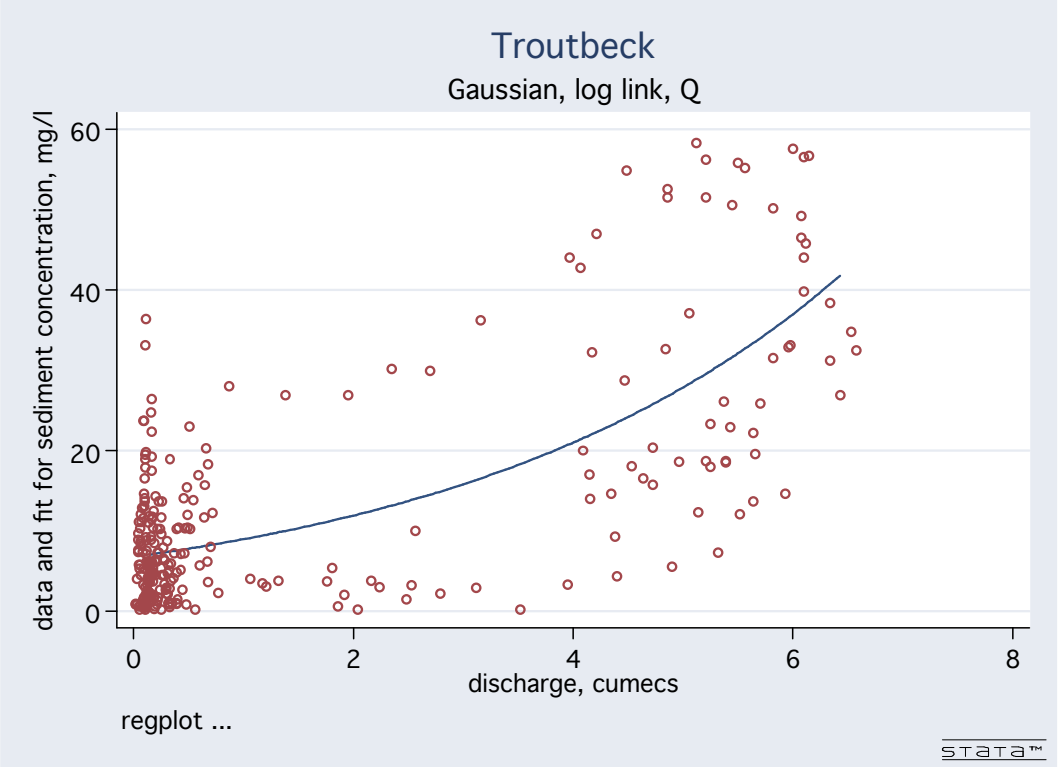

Troutbeck
sediment concentration, mg/l
0
10
20
30
40
0
2
4
6
discharge, cumecs
back transformed
smeared retransformation
log link, log Q, normal
log link, log Q, gamma
Stata™

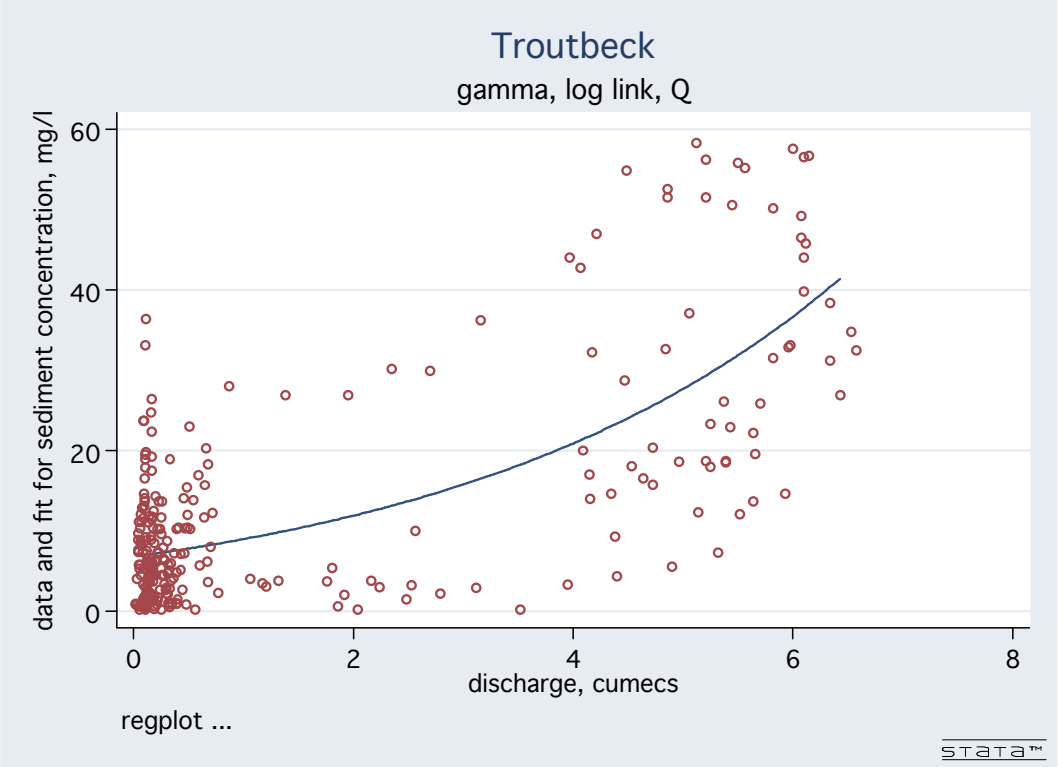

Troutbeck
Gaussian, log link
Response residuals
30
20
10
0
-10
-20
-4.074542-
-1.838851-
-.4462871-
ln Q
rdplot, g(3) ...
stata™

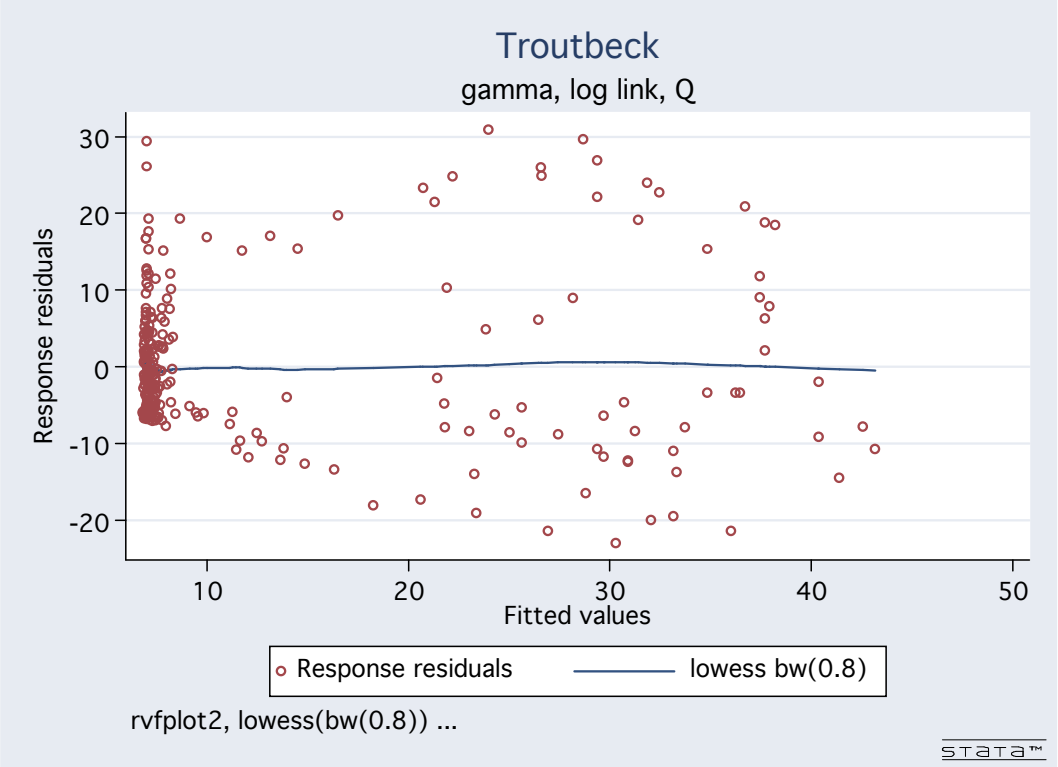

Troutbeck
gamma, log link
Response residuals
30
20
10
0
-10
-20
-4.074542-
-1.838851-
-.4462871-
ln Q
rdplot, g(3) ...
stata™

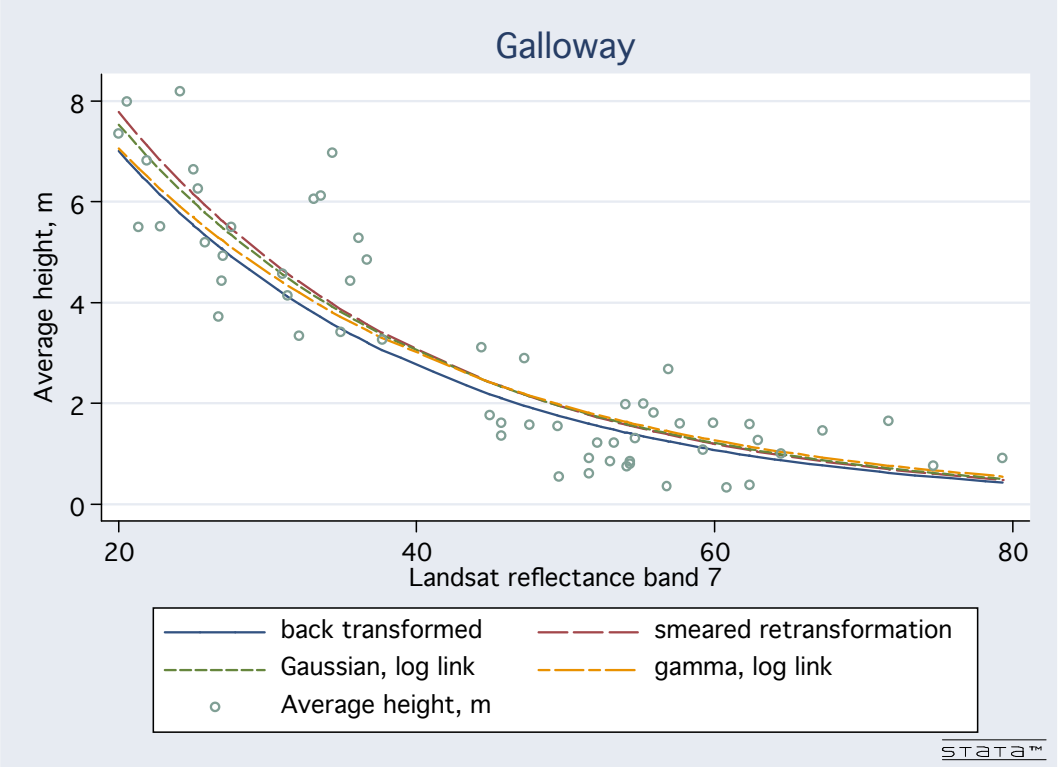

Troutbeck
gamma, log link
Response residuals
40
20
0
-20
-4.074542-
-1.838851-
-.4462871-
Graphs by ln Q
rdplot box, g(3) ...
stata™

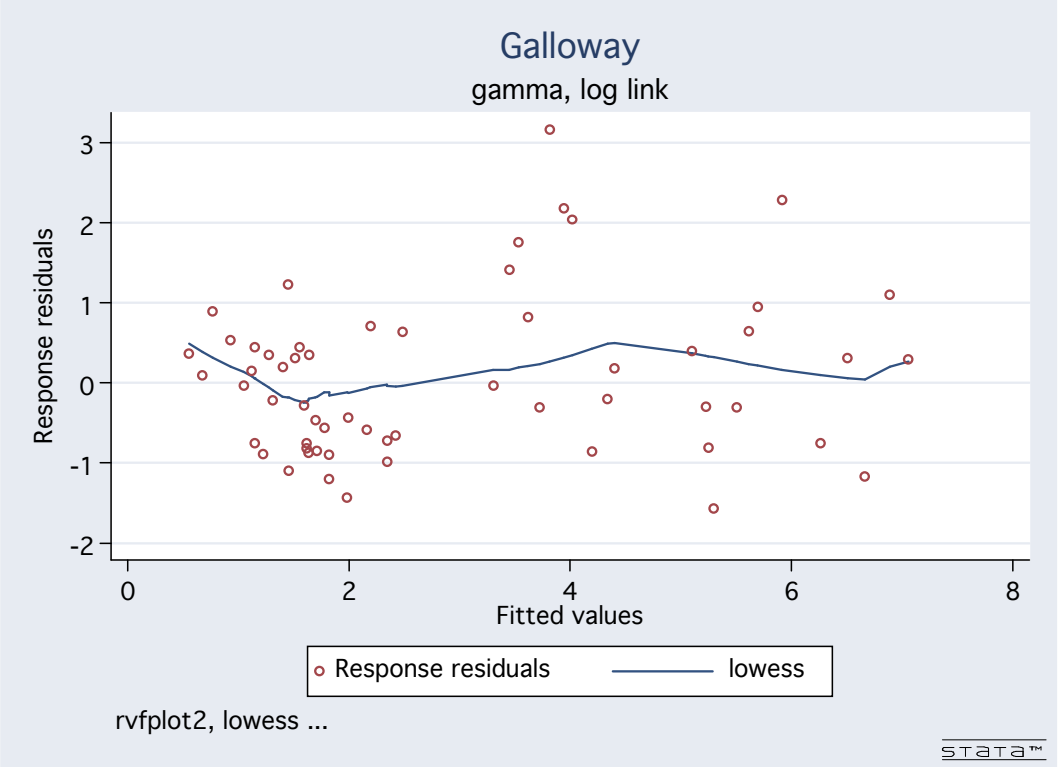

Troutbeck
gamma, log link, log Q
Response residuals
40
20
0
-20
5
10
15
20
25
Fitted values
Response residuals
lowess bw(0.8)
rvfplot2, lowess(bw(0.8)) ...
stata™

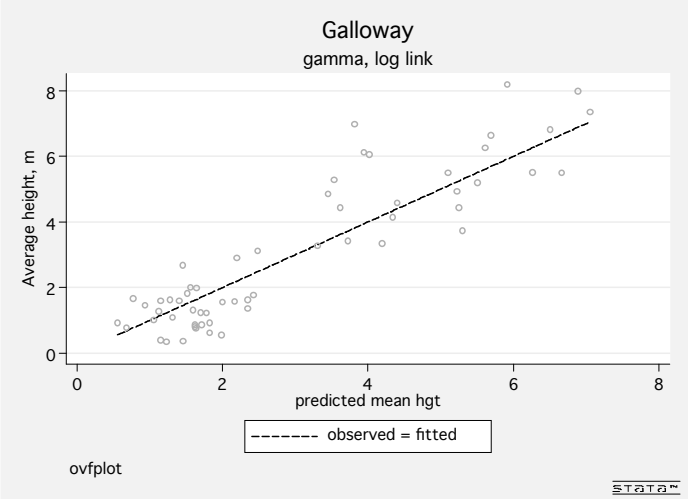

Troutbeck
Gaussian, log link, Q
data and fit for sediment concentration, mg/l
60
40
20
0
0
2
4
6
8
discharge, cumecs
regplot ...
stata™

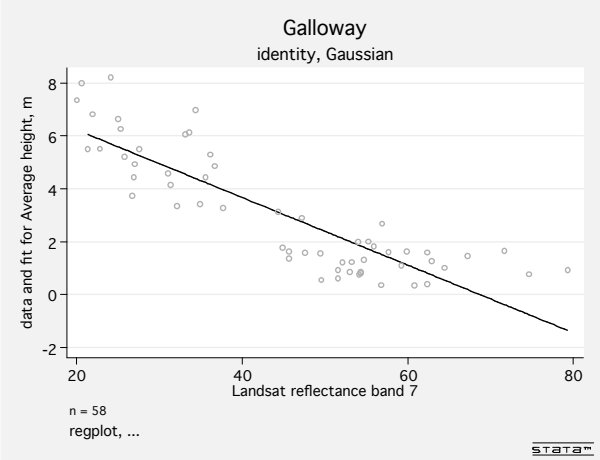

Troutbeck
gamma, log link, Q
data and fit for sediment concentration, mg/l
60
40
20
0
0
2
4
6
8
discharge, cumecs
regplot ...
stata™


Troutbeck
gamma, log link, Q
Response residuals
30
20
10
0
-10
-20
10
20
30
40
50
Fitted values
Response residuals
lowess bw(0.8)
rvfplot2, lowess(bw(0.8)) ...
stata™


Galloway
Average height, m
0
2
4
6
8
20
40
60
80
Landsat reflectance band 7
back transformed
smeared retransformation
Gaussian, log link
gamma, log link
Average height, m
stata™


Galloway
gamma, log link
Response residuals
3
2
1
0
-1
-2
0
2
4
6
8
Fitted values
Response residuals
lowess
rvfplot2, lowess ...
stata™


Galloway
gamma, log link
Average height, m
0
2
4
6
8
predicted mean hgt
observed = fitted
ovfplot


Galloway
identity, Gaussian
data and fit for Average height, m
8
6
4
2
0
-2
20
40
60
80
Landsat reflectance band 7
n = 58
regplot, ...
stata™